# Groq's LPU Accelerates Inference Without Governing It

by Nick Clark | Published March 28, 2026 | PDF

Groq developed the Language Processing Unit, custom silicon designed specifically for LLM inference that delivers tokens at speeds no GPU-based system can match. The deterministic execution model eliminates the scheduling overhead of GPU-based inference, producing consistent, ultra-low latency output. The hardware engineering is a genuine breakthrough in inference performance. But accelerating inference with custom silicon without adding semantic admissibility evaluation produces ungoverned output at unprecedented speed. The faster the hardware generates tokens, the more critical it becomes that each token is evaluated for semantic admissibility before commitment. Inference control provides this gate inside the generation loop, governing output at the speed the LPU delivers it.

---

## What Groq built

The LPU architecture uses a deterministic dataflow model that schedules compute at compile time rather than at runtime. This eliminates the dynamic scheduling overhead that makes GPU inference latency variable and slow. The result is inference that produces tokens at hundreds or thousands of tokens per second per user, with consistent latency that does not degrade under load. The GroqCloud API serves open-source models including Llama and Mixtral at speeds that enable previously impractical applications.

The hardware innovation is real. By rethinking inference from the silicon up, Groq achieves performance that software optimization on GPUs cannot match. The platform serves model output at hardware speed. The governance properties of that output are determined entirely by the model. The hardware accelerates delivery. It does not evaluate what it delivers.

## The gap between hardware speed and governed output

Groq's speed makes real-time AI applications feel instantaneous. Conversational agents respond as fast as humans can read. Code generation streams at writing speed. Content generation produces paragraphs in fractions of a second. These applications are precisely where governance matters most because the speed eliminates any human review opportunity between generation and delivery.

At GPU inference speeds, there is at least a temporal window where output could theoretically be intercepted and evaluated. At LPU speeds, the output reaches the consumer before a post-generation review system could react. The governance must be inside the generation loop because there is no meaningful after-generation window at hardware-accelerated speed.

## What inference control enables

The admissibility gate operates within the generation loop at the speed of the LPU's token production. The gate evaluates each candidate token or token group against persistent semantic state that is maintained in low-latency memory alongside the model state. The evaluation adds minimal overhead because it checks semantic consistency against pre-computed state rather than performing independent analysis of each token.

The entropy-bounded property constrains the semantic scope of generation at hardware speed. The pre-generation distinction ensures that admissibility is evaluated before each token is committed to the output stream, not after the full response is generated. The rollback-recovery mechanism provides governed recovery for the rare cases where a committed token sequence is discovered to be inadmissible mid-generation.

## The structural requirement

Groq's LPU represents a breakthrough in inference hardware performance. The structural gap is semantic governance at hardware speed: the admissibility evaluation inside the generation loop that ensures the fastest inference in the industry is also governed inference. Inference control as a computational primitive transforms hardware-accelerated generation into hardware-accelerated governed generation. Speed without governance produces ungoverned output faster. Speed with inference control produces governed output at a pace that transforms what AI applications can do.

Inference Control All 21 steps →

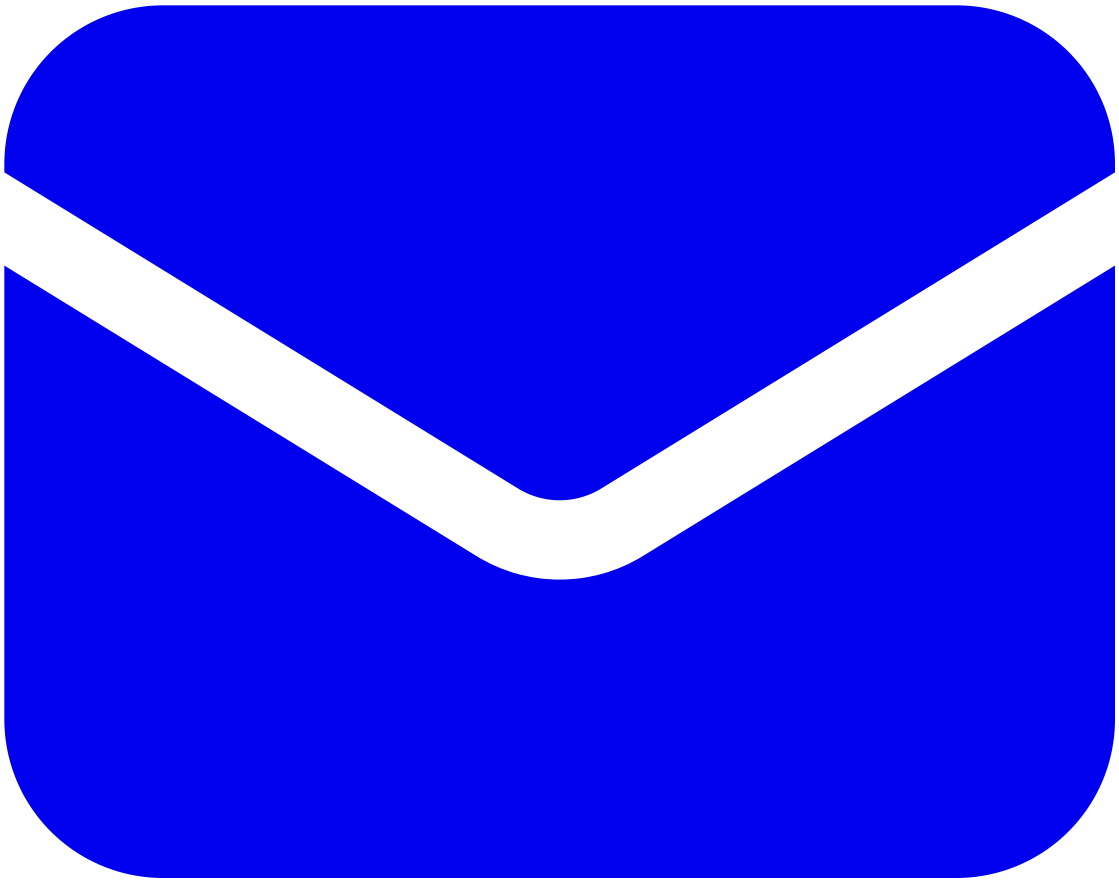
Govern inference at the point of generation.

Primary Technical Disclosure
○ Inference-Time Semantic Execution Control
Secondary Technical
○ Inference as Semantic Execution○ Semantic Admissibility Gate○ Entropy-Bounded Semantic Admissibility○ Inference-Time Semantic Budget○ Semantic Rollback and Checkpoint Recovery○ Multi-Model Arbitration With Shared Semantic State○ Structural Elegance Evaluation○ Rights-Grade Inference Governance○ Semantic State Object○ Semantic State Object Schema○ Inference Transition as Mutation○ Trust-Slope Continuity Across Inference○ Anchored Semantic Resolution○ Semantic Lineage Recording○ Policy-Governed Inference Execution○ Partial State Handling○ Model-Agnostic Inference Governance○ Pre-Generation vs Post-Generation Distinction○ Affect-Modulated Inference Admissibility○ Integrity-Aware Inference○ Confidence-Gated Inference Advancement○ Inference Deployment Embodiments
Applications (General)
○ Safety Without Alignment Theater: Why Structure Beats Supervision○ How Commercial AI Platforms Reduce Prompt Size, Drift, and Governance Risk at Scale○ When Execution Governance Becomes a Competitive Advantage — The Layer After LLM Gateways○ Enterprise LLM Governance at the Point of Generation○ Healthcare AI Admissibility Before Clinical Output○ Inference Control for Legal Document Generation○ Inference Control for Financial Advisory Output○ Inference Control for Education Content Generation○ Inference Control for Government Communications
Applications (Specific)
○ Einstein Generates Without Semantic Admissibility○ Databricks Serves Inference Without Semantic Gates○ Snowflake Cortex Generates Without Admissibility Gates○ Hugging Face Serves Models Without Semantic Governance○ Cohere's Enterprise LLM Has No Semantic Admissibility Gate○ Together AI Optimizes Inference Speed, Not Inference Governance○ SageMaker Serves Models Without Semantic Admissibility○ Vertex AI Generates Without Per-Transition Admissibility○ Azure ML Deploys Models Without Admissibility Gates○ Modal Runs Inference Fast Without Governing Output○ Replicate Serves Open Models Without Semantic Governance○ Fireworks AI Optimizes Speed Without Governing Semantics● Groq's LPU Accelerates Inference Without Governing It○ Cerebras Achieves Wafer-Scale Inference Without Semantic Governance
Inference Control overview →
AQ
deterministic
autonomy

and brand. Other names may be trademarks of their respective owners.



Last updated: 2026-03-03




- 
- nick@qu3ry.net
- 72 28 14 36 01

Invented by Nick Clark | Founding Investors: Devin Wilkie